Author's response to reviews

Title: Neuropsychological effects of chronic low-dose exposure to polychlorinated biphenyls (PCBs)

Authors:
   Martin Peper (peper@psychologie.uni-freiburg.de)
   Martin Klett (martin.klett@rhein-neckar-kreis.de)
   Rudolf Morgenstern (rudolf.morgenstern@charite.de)

Version: 2 Date: 8 August 2005

Author's response to reviews: see over
Re: Ms Martin Peper, Martin Klett, and Rudolf Morgenstern: Neuropsychological effects of chronic low-dose exposure to polychlorinated biphenyls (PCB)

Dear Dr. Grandjean,

Dear Dr. Ozonoff,

Thank you for your interest in our above-named manuscript. I am sorry for the delay it took to prepare the revision. This was due to the end of the summer term and to a job talk.

We have now revised the manuscript and prepared a point by point response to the comments from all reviewers.

Furthermore, the ms was revised according to the instructions for authors. Please feel free to apply “close editing” if necessary.

We hope that the revised version will now meet with your and the reviewers' approval.

Yours sincerely,

Martin Peper, PhD MD
encl ms
Response to Comments

Comments of D. Bellinger:

Given that this study is described as exploratory (since it is inappropriate not to detect subtle
differences at this stage of research), the Bonferoni correction for alpha renders this more of
a confirmatory study, increasing the likelihood of overlooking subtle associations. This risk is
increased, as well, by the very modest sample size.

We fully agree with D. Bellinger’s comment (and his opinion published in his 2004 paper in
which he criticizes the overly conservative techniques of traditional null hypothesis testing in
behavioral neurotoxicology). In fact, there is a large amount of literature criticizing null
hypothesis testing (NHT) and we have addressed this point clearly in the Methods section.
In principle, the notion of demonstration (or replication) of a certain effect can be interpreted
in two ways: (1) has the effect been replicated/demonstrated by the present data and (2) is this
effect in principle replicable or demonstrable. The first aspect is traditionally approached in
the context of NHT whereas the second refers to reporting estimation statistics (Greenwald et
al.).

In the traditional, distribution based NHT approach, sample size is the critical determinant of
whether an effect can be demonstrated. However, NHT does not provide information as to
whether an effect can be potentially replicated in larger study groups. Therefore, we applied a
dual approach in the present work: (1) The question related to the demonstration of a PCB
effect was answered by traditional NHT approach (Bonferoni correction for alpha; since this is
too conservative as recognized by the reviewer, a correction of the Bonferoni correction
was applied, see text). (2) The question of potential replicability independent of sample size
was answered by an estimation statistics approach supplementing the NHT statistics (e.g.,
Greenwald et al.). In addition to testing whether an effect is demonstrated/replicated with a
certain statistical significance, the effect size $d$ of the behavioral measure as a continuous
indicator of a toxic exposure effect was reported. Bellinger (2004) suggested that a useful
strategy to avoid the difficulties of a distribution-based approach is to express performance
differences in a metric-free form. The effect size $d$ represent such a metric-free measure.
An improved description of this rationale described above is now provided in the Methods
section. For further comment on sample size see response to Dr. Schantz below.

The effect sizes listed in Tables 3 and 4 frequently exceed 0.7 and range as high as 1.38 for
$d_2$. A value of $d<0.2$ was used as the criterion for low performance on a behavioral variable,
and inspection of the $ds$ in Table 4 shows that values for all but a few variables were this
extreme. (Although I did wonder how a difference of 0.2 in the group $T$ score means on WAIS
Similarities, for which the SD was 9 in both groups, could result in a $d_2$ of 1.38.)

Firstly, there appears to be some misunderstanding concerning the $d$-values: Indeed, many $d_2$
effects were GREATER than +0.2 because performance of both teacher groups was clearly
above-average (in relation to the normative samples of the tests).

Secondly, according to our count, 46 percent of $d_1$ results fell below at least $d_1<-0.2$ (which
is not “all variables”). A value of $d<0.2$ was not used as a criterion for low performance on a
behavioral variable but as an indicator of at least moderate replicability. Effect sizes lower
than 0.2 are not represented in the Tables because they are not meaningful in terms of
replicability.

In the similarities subtest, both groups did not show substantial differences in verbal
intelligence (that is, a $d_1 < 0.2$). However, in comparison to the distribution of normative
values (mean 50, SD 10), both groups were 1 SD above average. Thus, the $d_2$ of 1.38 (as well
as other verbal measures) simply represents the above average performance of teachers
populations (mean $T$ in both groups: 63). The fact that both teacher’s groups showed
normative values and effects sizes suggesting a greater level of verbal intelligence is briefly mentioned in the Results section.

Furthermore, some of the effect sizes for the associations that did not meet the corrected alpha level were substantial, considerably larger than those typically observed in studies of low-dose chemical exposures. As indicated in the paragraph Dose-response relationships in the Results, some of the correlations observed between PCB28 and behavioral variables were as high as 0.54, with several others >0.3. These are extremely high correlations for multi-determined, complex outcomes such as performance on a neuropsychological test. This comment refers to five correlation coefficients with PCB28 which were not statistically significant according to NHT but showed an elevated association. We tried to evaluate whether the size of these correlations was artefactual as suggested by the reviewer. However, as stated in the ms, these correlations remained relatively stable even when nonparametrical reanalyses were done or when other variance components were controlled. Therefore, we see no other way of further reanalysing these data.

The alpha correction used leads the authors to minimize the potential importance of these correlations and effect sizes, as does their use of the criterion of clinical impairment at the individual level to judge the importance of the associations (statements such as, [differences] which, however, were within the range of the normative population and differences between exposed subjects and non-exposed controls were relatively small and because mean normative values did not exceed the levels required for the identification of impaired persons.

Applying an individual criterion for impairment to the interpretation of group differences on a population basis can be highly misleading, as I argued in a recent paper in Environmental Research (2004) entitled "What is an adverse effect".

We did not use alpha correction “to minimize the potential importance of these correlations and effect sizes”. This is not true because the dual analysis approach as described above was applied from the beginning. This aimed not only at the demonstration of an effect (i.e., testing a global null hypothesis) but also at evaluating the demonstrability (i.e., potential replicability in larger groups). This dual approach is clearly explicated in the original ms. Because there seems to be misunderstanding, however, a section was added in the Methods section that better explains this approach. The Results section was also improved.

The reviewer also mentions that applying a criterion of clinical impairment at the individual level can be misleading. The statement “required for the identification of impaired persons” was certainly not appropriate since it implies that cut off values or a clinical criterion was used in the present study. However, this was not the case. In fact, the “d2” effect size was meant indicating the deviation from the expected distribution of the normative sample. The statement was deleted and the section was rewritten.

Specific Comments
1. What considerations resulted in the decision to have a sample size of 60? Were any selection criteria (inclusion/exclusion) applied? Was the number limited by the number of teachers working at the exposed school?

The number of PCB-exposed subjects (N=30) was limited by the number of persons working at the PCB-contaminated school (which can, for example, be verified by a view on the employees page of school’s web page); the 30 non-exposed teachers were matched by age and education. The fact that the total group was examined has been added in the Study population section.
Reference is made at the end of the Discussion that the associations, might have been moderated by potential confounders such as differences between city and rural populations in lifestyle and nutritional habits. The Methods indicates only that teachers were drawn from two high schools near Heidelberg. More information needs to be provided about these potential differences.

PCB-exposed teachers were drawn from a rural region secondary high school (“Realschule”) (30 km distance from Heidelberg) and non-exposed teachers from another secondary high school located in the city area of Heidelberg. This is now better described in the Study population section. Potential differences are further discussed below.

The statement that the group differences might have been due to different performance profiles associated with school type, or motivational factors needs to be explained more. Why would performance profile differ by school (in the absence of an effect of PCB exposure), and what motivational factors might have varied between schools (and why)?

Firstly, to expect school type differences (in the absence of an effect of PCB exposure) which was mentioned at the end of the Discussion is probably too speculative. Both schools were secondary modern schools (“Realschule”). Although some of these schools might differ with respect to their educational programs, it is almost impossible to describe and quantify these potential differences. Since no difference of school type can be detected, this was deleted in the Discussion.

Secondly, differences between city and rural populations in lifestyle and nutritional habits: The used term “Life style” was probably too broad. Differences of both groups (which also include differences of city and rural populations) were assessed as nutritional habits, clothing, use of wooden interiors potentially treated with preservatives etc. These factors were assessed by means of the comprehensive questionnaire. These results were condensed in the section “External exposure”. In our opinion, it is not necessary to further go into details here.

Thirdly, performance profile related to motivational factors: There were no performance differences concerning teachers from rural region and city teachers. Motivational factors were different because of the increased awareness of PCB-exposed teachers (rural high school) versus the non-exposed teachers (city high school). The hypothesis that differences related to prior information about exposure status could be related to elevated complaints or to changes of behavioral performance was tested by comparing subjects with low objective PCB exposure with matched controls. Therefore, 10 subjects with low PCB were identified (PCB \( \leq 0.1 \) µg/l) and compared with matched controls. Results were not significant on a corrected level and the trends are briefly described in a separate section in the text (new section “Re-analysis of subgroup with low PCB 28 blood levels” in Results). However, it is mentioned in the Discussion that these results must be treated with caution because current (low) blood values might not appropriately reflect the prior long-term exposure to PCBs.

2. Were the neuropsychological evaluations conducted in a blinded fashion or did the tester know the school at which an examinee worked?

Because evaluations were done near or in the respective schools, the examiner was not blind to the exposure site; however, they were blind with respect to the objective exposure status. The effects of exposure awareness are discussed above.

3. The Ms indicates that no teacher had a history of neurological or psychiatric disorder. It seems unlikely that, in a sample of 60 adults, none had a history of, in particular, a psychiatric disorder.

The typical incidence of psychiatric disorders ranges from to 1 to 5 percent of the population in Germany. In our teachers, no history of productive psychiatric diseases was found. This can be explained by the following reasons:
(1) Statistically, we would expect 0.3 - 1.5 persons per 30 teachers with a psychiatric disorder. Due to the low number expected to have a psychiatric disease, it is not unlikely that the two different cohorts were void of teachers with a history of psychiatric disorders.
(2) As our teacher’s age range was between 40 and 60 years, psychiatric disorders can be ruled out to a large degree since most active psychiatric diseases will appear prior to the age of 40.
(3) As a consequence, persons with psychiatric diseases can and will not be employed as teachers any further. Even if a single subject had such a diagnosis, this would not have biased our data considerably.

How was this information collected, and what criteria were applied?
As indicated in the Methods section, information about psychiatric disorder history was collected in a medical screening examination; criteria were assessed by a clinical interview and additional self-report by questionnaire. This is more clearly described now in the Methods section, but see also comments above.

4. Additional information about the measurement of the concentrations of PCB in air samples would be helpful. How were the samples collected? Were they 24-hour samples? When were they collected in relation to the participants’ blood PCB levels?
PCB concentrations in air samples have been collected during 24 h periods with closed doors and closed windows and temperatures of 20-22°C. Blood samples were collected during the medical examination which was followed by the neuropsychological test session. The interval between blood sampling and psychological testing was 1 to 3 days. The interval between last exposure to contaminated indoor air and blood sampling was four weeks. The average interval between first air sampling and blood sampling was 8 weeks. This additional information is now given in the respective Methods section.

Near the end of the section, External Exposure, please clarify the statement that an accumulation of potential risk factors could not be confirmed for either group.
The statement which refers to the vocational risk factors mentioned in the sentences before was clarified: “The total frequency of these potential vocational risk factors was not significantly different.”

5. What was the temporal relationship between blood sampling and potential PCB exposure in the school? Given that exposure appeared to consist primarily of low chlorinated congeners with shorter half-lives, some exposure misclassification could have occurred if exposure and blood sampling were separated by a few months.
The temporal relationship between last PCB exposure and blood sampling was four weeks (see comment above).

Also, I am not clear on the implications of the fact that the GC-ECD was replaced by GC-MS. Why not just say that the samples were measured by GC-MS, if those are the levels used in the analyses?
The authors agree that GC-MS was the correct method to measure and characterise PCB-congeners. Parallel analyses of the samples done with the GC-MS and GC-ECD methods did not show differences (as confirmed by the laboratory of the Landesgesundheitsamt Baden-Württemberg, Stuttgart). The respective sentences in the Biological Monitoring section were simplified.
6. The MS indicates that a total toxicity index was not calculated, since complete information about all congeners was not available. No study ever measures all PCB congeners, yet TEQs are estimated, so this cannot be a sufficient reason not to attempt to do so here. We were mainly interested in health effects related to the neurotoxic potency of the relatively low indoor-air PCB-pattern in a public building, since a number of teachers were complaining about fatigue and negative effects on concentration and memory. The TEQ-philosophy was not used on purpose since TEQs concern mainly the dioxin-like effects. Since our study did not focus on coplanar PCBs, this effect was of little interest. The last sentences at the end of the Biological Monitoring section/Methods was changed.

7. What is the basis for the statement that computerized testing was added to increase the sensitivity of the battery in the area of executive and attentional functions? Why is this statement made specifically with reference to executive and attentional functions as opposed to other neuropsychological functions? What evidence can be marshaled to indicate that computerized testing is, in fact, more sensitive than examiner-administered tests for these domains?

Certain performance domains can, in fact, be more reliably measured by computerized testing (reaction time, omission errors in attention tasks, see Hartman, 1995) than examiner-administered tests. We think that this is quite obvious. Nevertheless, we deleted the critical statement and improved the TAP-description.

8. How the preceding factor analyses of control subjects data were conducted and how it structured the analyses is not clear. These analyses are not mentioned in the Results section. A new section was added in Methods/Neurobehavioral assessment to explain FA methods: “…Since neuropsychological measures are partly intercorrelated, explorative factor analyses (principal component analyses with varimax rotation, using data from a pool of available control subjects, N=72) was computed separately for behavioral and self-report variables. The scree-test suggested 8 factors for behavioral measures (each explaining 8 to 15 percent of the variance, total 80 percent) and 5 factors for self-report measures (each explaining 15 to 30 percent of the variance, total 70 percent). The obtained factor structure was used to group the scores and to derive factor descriptions (using variables with loads >.50). Moreover, median effect sizes were computed for each factor and presented in the Tables.”

In addition, it would of course be possible to present Eigenvalues and Factor loads (see original German text below) but we think that this would be too much of irrelevant information.

We think that the results of the factor analyses should not be presented in the Results section, as suggested by the reviewer, but in the Method section because (1) it is derived from a larger pool of control subjects and (2) it serves to pre-structure the Results section and facilitate the interpretation of findings.

[Original German text presenting Eigenvalues and Factor loads: Faktor I (Eigenwert \(\lambda_1=3.6\); relativer Varianzanteil: 15.2%) kann als fluide verbale Intelligenz bezeichnet werden, mit Schwerpunkt in verbalen Benennungs- und Abstraktionsleistungen, die den Markrievtablen Benennungsgeschwindigkeit (FWIT) (\(a_{ij}=.84\), Lesegeschwindigkeit (.74), Interferenzanfälligkeit beim Lesen (.74) sowie Zahlenmerkspanne vorwärts (.74) und Gemeinsamenkeitenfinden (.52) entsprechen.

Faktor II (Eigenwert \(\lambda_2=3.2\); relativer Varianzanteil: 13.5%) umfasst visuo-kognitives Gestalterkennen mit Identifikationsleistungen bei fragmentierten Wörtern (.91) und Bildern (.80), beide Variablen wurden aufgrund ihrer Redundanz zusammengefasst.

Faktor III (Eigenwert \(\lambda_3=2.9\); relativer Varianzanteil: 12.3%) entspricht visuo-motorischen Leistungen mit Mosaiktest (HAWIE) (.81) und der visuellen Arbeitsgedächtnisaufgabe mit motorischer Interferenz (.86).

Faktor IV (Eigenwert \(\lambda_4=2.5\); relativer Varianzanteil: 10.7%) entspricht der krittalisitseren verbalen Intelligenz mit Allgemeinwissen (HAWIE) (.73) als Markrievabla. Auch der HAWIE-Gesamt-IQ lädt auf diesem Faktor (.53).
Faktor V (Eigenwert \( \epsilon_5 = 2.4 \); relativer Varianzanteil: 10.4\%) bezieht sich auf Arbeitsgedächtnisfunktionen mit visueller Merkspanne vorwärts und rückwärts (.87) und Zahlenmerkspanne rückwärts (.69) sowie ferner den Auslassungsfehlern in den TAP Untertests Arbeitsgedächtnis, Reaktionswechsel und geteilte Aufmerksamkeit.

Faktor VI (Eigenwert \( \epsilon_6 = 2.3 \); relativer Varianzanteil: 10.2\%) entspricht der Dimension Selektive Aufmerksamkeit und Alertness mit den Markiervariablen Trail-Making-Test A und B (.81), dem TAP-Kennwert Phasische Alertness (.62) sowie ferner dem Aufmerksamkeits-Belastungstest.

Faktor VII (Eigenwert \( \epsilon_7 = 2.3 \); relativer Varianzanteil: 10.1\%) entspricht der Dimension Gedächtnis, mit den Markiervariablen Wiedererkennen fragmentierter Wörter (.78), Logisches Gedächtnis (WMS-R) (.50) sowie ferner Wiedererkennen fragmentierter Bilder, Visuelle Reproduktionen (WMS-R) und Recurring Figures.

Faktor VIII (Eigenwert \( \epsilon_8 = 1.7 \); relativer Varianzanteil: 7.9\%) entspricht Speziellen Funktionen des präfrontalen Kortex mit den Variablen Verbale Flüssigkeit (formallexikalische Wortflüssigkeit/divergentes Denken) (.86) und Prozedurales Lernen (TvH, .63).

Eine weitere Faktorenanalyse, welche die zusätzlichen Variablen (z.B. TAP) der Untersuchung 1 einschloss, bestätigte diese Struktur, wobei die Reaktionszeiten in den komplexen Aufmerksamkeitstests einen eigenen Faktor Psychomotorische Geschwindigkeit und Reaktionsbereitschaft in komplexen Wahlreaktionaufgaben konstituierten.

Self-report variables:

Faktor I (Eigenwert \( \epsilon_1 = 6.8 \); relativer Varianzanteil: 29.6\%) kann entsprechend der Polung der Einzelskalen als Dimension eines allgemein verzerrten körperlichen und emotionalen Wohlbefindens bzw. im Sinne der Persönlichkeitseigenschaft Emotionalität interpretiert werden. Als Markiervariablen können die Skalen S I, S II, 1, 5, 7, 8, 9 und N des Freiburger Persönlichkeitsinventars FPI-R (durchschn. \( \alpha > .60 \)), die Beschwerdenliste B-L (.58) und das Beck-Depressions-Inventar (.54) dienen. Darüber hinaus sind auch die verschiedenen Symptomskalen 1, 2, 3, 4, 6, 7, 8, 10, 11 der Freiburger Beschwerdenliste FBL-R mit diesem Faktor verbunden.

Faktor II (Eigenwert \( \epsilon_2 = 4.4 \); relativer Varianzanteil: 20.3\%) umfasst den aktuellen emotionalen Befindlichkeitszustand im Sinne eines State-Faktors. Dieser wird repräsentiert durch die aus der Eigenschaftswörterliste gewonnenen Variablen Deprimiertheit/State-Angst (.81), Allgemeine Desaktiviertheit (.81), Emotionale Gereiztheit (.80), geringes allgemeines Wohlbefinden (.72) sowie geringe leistungsbezogene Aktivität (.53).

Faktor III (Eigenwert \( \epsilon_3 = 1.6 \); relativer Varianzanteil: 14.9\%) bezeichnet die Persönlichkeitsdimension der Extra- vs. Introversie, wobei entsprechend der Polung die Ausprägung in Introversion (FPI-R E; .82), geringe Aggressivität (FPI-R 6; .80), Verschlossenheit/geringe Offenheit (FPI-R 10; .71) und geringe Leistungsorientierung (FPI-R 3; .56) als Markiervariablen dienen können.

Faktor IV (Eigenwert \( \epsilon_4 = 2.2 \); relativer Varianzanteil: 16.4\%) entspricht der Persönlichkeitsdimension Sozialität mit den Variablen Gehemmtheit (FPI-R 4; .64) und geringe soziale Orientierung (FPI-R 2; .63). Fremdeinschätzungen in den Bereichen emotionale Auffälligkeiten (.88) bzw. kognitive Beeinträchtigungen (.60) laden ebenfalls hoch auf dem Faktor Sozialität. Auch die Skalen FBL-R 9 (Unruhe/Motorik) und 5 (Kopf-Hals-Reizsyndrom) sind mit diesem Faktor assoziiert.

Faktor V (Eigenwert \( \epsilon_5 = 2.5 \); relativer Varianzanteil: 18.8\%) entspricht dem aktuellen Leistungszustand im Bereich Antrieb und Aufmerksamkeit. Markiervariablen sind Antrieb (.89), Ermüdbarkeit (.83) und Ablenkbarkeit (.48) des PEDA.

9. Why gender received so much attention as a potential confounder is puzzling, especially since it was unrelated to PCB exposure variables and behavioral outcomes. Given the small size, the degree of freedom used up by adjustment for gender might have been used more productively adjusting for some other factor that actually functioned as a confounder in this data set.

It is surprising that the reviewer criticizes the use of gender as a covariate in the light of the many findings suggesting that gender is an important source of variance and major confounder in neuropsychological studies as well as exposure assessment. It has also been discussed with respect to PCB exposure (e.g., Schantz, Widholm & Rice, 2003). If this variance component cannot be controlled by matching it should be included as an additional factor. Even when the portion of the variance explained by gender is relatively small, we should control for the gender main effect (in the absence of significant exposure x gender effects) to obtain and report unbiased PCB effects. The majority of readers would certainly not conform to the opinion that we should report effects that are biased by gender to a certain degree. In any case, the additional degree of freedom would not substantially change the present results and their interpretation.
This section does indicate that, in separate analyses, adjustments were made for openness and alcohol consumption on self-report and estimated intelligence level and alcohol on behavioral measures. Why these variables and why only for these outcomes? How were they identified as potential confounders. Table 1 shows that the two groups did not differ in IQ or alcohol (although the latter variable was clearly highly skewed).

This statement of the reviewer refers to details of the correlation analysis (last section of Methods and Results/Dose-response-relationships). Self-reported complaints and mood state showed no substantial positive association with PCB or the cumulative index (see Results). However, openness differed between groups (PCB subjects were more reserved in answering the questionnaires, p=.06, see Table 3) and alcohol consumption showed a moderate effect size (n.s., but d=0.30).

Even in the absence of strong confounding effects, it is reasonable to control for low to moderate effects in order to avoid spurious correlations and suppression effects. Thus, the correlation analyses were done with and without these confounders. Since confounding effects on self-report measures were relatively small, we think that that it is not necessary to do a major revision of the respective section. However, the last section in the Methods/Statistical Analysis Section was adapted.

Estimated intelligence (as a measure of exposure-independent general intellectual ability level) has generally been recognized as a major confounding source in behavioral neurotoxicology studies. Controlling for estimated intelligence includes the effects of potential demographic confounders related to intelligence. Similarly, consumption of alcohol may have a direct effect on behavioral measures independent of exposure. This variable has frequently been used as a predictor in regression analyses. This is also more clearly expressed in the Methods section. Thus, correlations with behavioral variables were computed with and without consideration of these confounders. IQ and alcohol were partialled out in order to control for and to avoid spurious correlations. This is an appropriate procedure, in particular, when common regression techniques cannot be applied due to small N.

Given the lack of confounding effects, our statement in the Results/Dose-response-relationships section that correlations “survived a reanalysis with rank correlations and partialing out of estimated intelligence level and alcohol consumption” is appropriate and shows that the authors have made efforts to control for typical confounding variables.

Finally, it is quite trivial that intelligence/performance and personality factors such as “Big five” represent relatively independent psychological domains. Therefore, it is hard to conceive of a reasonable hypothesis of how self-report outcomes could be confounded by general intelligence. Conversely, it is hard to conceive of reasonable hypotheses of how behavioral outcomes could be confounded by openness as assessed by questionnaire.

In other words: it is necessary to consider different confounders for self-report and for behavioral measures. This is also briefly mentioned in the Methods section.

10. Although the Data analysis section indicates that a corrected alpha of 0.004 was applied, the Results refers to significant associations for which P-values were >0.004. Also, given the study design (matching exposed and controls on various factors), were the analyses conducted using methods appropriate for matched data?

The Table footnote clearly explicates the symbols: Table 4: 2 F-value for group effect adjusted for gender; *: significant relative to original-alpha; ***: significant relative to Bonferroni-adjusted alpha. The statement “Both corrected and uncorrected α’s are given in the Tables” was added to the Statistics Section. The second question refers to the problem of whether matching may lead to dependent data which should be analysed by, for example, repeated measurement-ANOVAs. I have repeatedly discussed this with the reviewers of previous papers and opinions vary. In the present study, however, matching was possible only with respect to some factors, other factors such as gender could not be considered. In our
opinion, it is not justified to talk of dependent samples. Using dependent statistics may yield slightly lower p-values. We also carried out such analyses but a greater effect (in favour of the H$_1$) was not observed.

11. Greenwald et al. is now included in the References section.

12. It isn’t clear what is being shown in Figure 2. Are these values only for PCB 138, 153, and 180, as the legend suggests? If so, this doesn’t seem very helpful insofar as the school-related exposure was determined to consist mostly of congeners 28 and 101. If PCB28 levels are reflected in Figure 2, how were the large number (>90%) of controls with PCB28 below the detection threshold represented?

Figure 2 was simplified and rearranged. The fact that the median of PCB28 control levels was below the detection threshold was noted in the figure legend.

13. In the first paragraph of the section, Neuropsychological results, I think the reference to Table 1 should be Table 3?

No, this is correct (reference to self-reported EQ complaints in Table 1). However, a redundancy of EQ scores in both tables was eliminated and the reference to the Tables was improved.

Similarly in the next paragraph in this section, should reference to Table 3 instead be Table 4?

Corrected.

14. In Tables 3 and 4, a delta is used to indicate a potentially relevant effect and refers the reader to the text for an explanation, but I could not find such an explanation. The explanation has been given in the footnote. In addition, it is mentioned now in the statistical analysis section as well as in the List of Abbreviations.

15. The self-report measures that differed between groups defined on the basis of external exposure (well-being, distractibility of mental processes, introversion) seemed to differ from those that were associated with biological measures of PCB (tiredness and slowing, emotional reactions). Can the authors comment on potential explanations for this? Does it suggest that the associations might be the result of chance?

We don’t believe that the differences between total group and subgroup self-report measures are a result of chance (see below). The total group described itself as more introverted and less aggressive. In contrast, the subgroup included subjects which described greater distractibility, inattention and slowing as well as aggressiveness. These differences in self-reports are all compatible with the notion of a slight hypofrontality. However, it is questionable whether this could be readily stated in the manuscript without overinterpreting the data. We added the statement that relative to the total PCB group, the PCB exposure subgroup described greater inattention, tiredness, distractibility as well as slightly more emotional and aggressive reactions. We also compared subjects without elevated blood values with the controls. The results are now given in a new section (see comment above).

16. The last sentence of Results states that differences were found on TAP phasic alertness and response shifting, yet the P-values are listed only as n.s. Why is attention drawn to these non-significant results and not to others?

Attention is drawn to the TAP results because they showed elevated effect sizes corresponding to hypothesis. This is clarified by adding "effect size". Since this section
focuses on the effect sizes, the given p-values were deleted to make this section more consistent.

17. The paragraph in the Discussion on the generalizations permitted (last paragraph in the section entitled External and internal PCB exposure) is quite confusing. First, it says that generalization does not seem warranted. Then, at least partial generalizations to other PCB exposure conditions might be possible. Then the last sentence of this paragraph suggests that generalization might not be possible because, a diverging PCB pattern or additional pollutants at another site may thus produce a different behavioral effect.
This section was rewritten and made more consistent.

Comments of B. Heinzow:

In this paper, the importance of PCB blood determinations following potential exposure is demonstrated, and a congener specific application is used to assess the role of specific PCB congeners or classes of congeners in mediating neurotoxic outcomes. From a statistical view, attempts to determine the role of individual congeners in mediating outcomes are corroborated by the fact that concentrations of individual congeners are highly correlated with each other and with total PCBs. In this study the unique role of congener 28 and 52 is shown, due to the fact that specific and unusual air-borne exposure has been identified that bears a different profile in comparison to food-borne PCB pattern.
PCBs are mixtures of multiple congeners, and different congeners may have very different actions but their relative potency to produce nervous system effects is unknown. From experimental work exposure to a single, relatively high dose of PCBs decreases the content of several brain neurotransmitters, while repeated exposure to lower PCB doses appears to affect brain DA metabolism, and can alter a number of physiological processes that may be important for developmental toxicity. This study addresses adults with a long-term exposure history and a different mechanism of neurotoxicity as in the developing brain must be considered. The study by Peper et al is well written and has used a wide array of current neurotoxicity testing tiers and hence produced multiple results and deserves publication to stimulate further research.
However as the authors point out, the power of the study is limited and findings should only be seen as preliminary and speculative. It would be interesting to continue with this type of study in children (exposed in schools and with a larger sample) and heavily exposed workers.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
1. In the tables the BMI is used in an unfamiliar manner [T-value], I am curious if this indicates a state of severe obesity in the participants or if it is a term referred to as the 'T value' to determine the normal body weight for given height, thus a T value gives the ideal body weight in kilograms, and is equal to height excess to 1 metre in centimetres. Most often the Body mass index (BMI) is a measure of body fat based on height and weight that applies to both adult men and women and expressed as : BMI [kg/m²]. I would like to see how does this compare to the study data? Where I calculated from table 1 for the exposed group 25 and for the controls 23, thus indicating no difference an no obesity.
The BMI values are now expressed in kg/m².
2. The last statement in Results that "differences were found in behavioral variables" would be more appropriate worded as no statistical differences were found ... Unless I interpret the values wrongly (given in parenthesis).

The statistical significance was omitted in this section and the section was rewritten (see comments to reviewer 1).

Discretionary Revisions (which the author can choose to ignore)
A noteworthy finding is the fact that the exposed group, especially the subgroup of 16 (12+4) displayed a significantly elevated PCB-exposure not only for PCB 28 and 52, what is expected, but also for the highly chlorinated congeners, which do not occur in indoor air. This might point to differences in metabolism or distribution, or is just coincidence. Please comment.

Firstly, the PCB-exposure was not significantly elevated for PCB 138, 153 and 180 in the total group. The effect sizes for these variables were relatively small (0.15-0.30). In the total group, PCB 138, 153, 180 and PCB sum were also correlated with age (r=0.27). This and other correlations from Table 2 suggests that additional or background exposure sources (e.g., leather clothing, vocational sources) might have been responsible for these associations.

Secondly, the differences were more distinct in the subgroup analysis, especially for PCB 180. In this group, duration of exposure (r=0.46) but not age (r=0.05) predicted PCB 180 values.

PCB 180 is a nonplanar congener which was also observed in the contaminated air (Figure 1). High chlorinated PCB-congeners from indoor origin might be responsible for greater blood levels because chronically exposed persons metabolize low chlorinated PCB’s faster than high chlorinated congeners which are metabolized slowly. Thus, the described situation is typically for chronically low dose PCB exposure.

On the other hand, however, PCB 180 was not elevated in comparison to age-related expected values as derived from a national sample (Figure 2). The significances of the PCB subgroup might thus be a result of low local incorporation as well as of the comparison with a control group with a relatively low background exposure. These aspects are briefly discussed in the ms.

Comments of S. Schantz:

This manuscript presents findings relating to the neuropsychological impact of exposure to PCBs from indoor air. The issue is an important one that has received little attention in the literature. The assessment procedures used to measure neuropsychological function are appropriate. However, there are a number of problems with the manuscript and the study. These include a lack of statistical power due to a very small sample size, the relatively low level of PCB exposure in the sample that was studied, the limited analysis of just a few marker PCB congeners, a questionable statistical approach, and over interpretation of results that were either not significant or only marginally significant. The PCB literature is also rather selectively cited in the introduction, which could lead to some false perceptions for readers who do not know the field.

The introduction needs to be better organized and better referenced. It would be better to cite review articles rather than selectively referencing a subset of publications from a subset of the individual studies. Also, several statements are misleading or incorrect. The reference to PCB 105, 118 and 156 as the most frequently detected congeners in the US population is not correct. These may be the most frequently detected congeners with AhR activity, but they are not the most frequently detected over all. A statement like this should be based on an overview
of the literature, not one isolated article. A summary of various studies would show that PCB 138, 153 and 180 are the most frequently detected and are also present in the highest amounts.

The introduction was rewritten and references were added. In particular, the section focusing on developmental effects was abbreviated. The statement concerning most frequent congeners was modified.

It is also misleading to say that neurodevelopmental effects are noted primarily in the striatum, prefrontal cortex and cerebellum. This may be due to a lack of studies on other brain regions rather than a lack of effects. Many brain areas have not yet been studied for PCB effects.

The corresponding sentence was improved.

The small sample size in this study together with the relatively low PCB exposure from air is a big concern. As a result, the study is seriously lacking in power to detect an effect. As the authors themselves note a sample size of 200-500 PCB-exposed individuals would have been needed and this study included only 30 exposed subjects. It is hard to draw any useful conclusions from such an under-powered study.

The study is clearly characterized as exploratory. The problems associated with sample size are clearly identified and questions related to power are discussed in the Statistics section and answered by means of a dual NHT/estimation statistics approach (see comments to reviewer 1). The reference to the 200-500 subjects was made to characterize the limitations of the traditional NHT testing approach. Since both approaches have already been implemented in the statistical framework of the original manuscript to a satisfying degree, principal changes or re-analyses were not necessary. However, the respective sections describing this approach have been improved and the Discussion has been modified to account for these limitations.

It should be mentioned that - as a consequence of the reviewer’s comment concerning sample size - clinical case studies as well as the vast majority of neuropsychological/behavioral neuroscience studies would have to be withdrawn from the publication process. In the light of the lack of published reports, this certainly seems exaggerated and does not correspond to the reality of neurobehavioral research.

It is certainly true that exposure by ingestion might produce larger effects than by inhalation. However, the consequence NOT to study groups that have been chronically exposed to certain air-borne PCBs in the low dose range would certainly be a serious mistake.

By the way, the PCB 28 exposure effect was large when compared to the present controls. Moreover, due to the interval of 4 weeks between termination of exposure and blood sampling, PCB 28 levels may have been underestimated. The latter statements were added in the “External and internal PCB exposure” Section/Discussion.

Another limitation is the decision to analyze only a few? marker? PCB congeners. It is unclear what percentage of the total PCBs in air these congeners typically represent or whether they are actually good markers that can be used to accurately estimate total airborne PCB exposure.

The present work is not a methodological contribution to PCB measurement! As explicated in the ms., the toxicology part of the current work has already been published by the state’s health authority (e.g., Gabrio et al. 2000; Schwenk et al., 2002). The relevant and new part is the neuropsychological assessment and it’s relation to internal exposure.

The investigated congener pattern was selected because it is a typical marker for polymer plasticisers that have been used in Germany in the 1970’s. This has been established by the responsible local authorities as a routine method for an estimation of air-borne PCBs in
Germany. We think that a replication of this discussion what actually represents a good marker is not necessary here. The reader is directed to the above references. Moreover, the Biological Monitoring Section was rewritten.

The authors state that more extensive analyses were not feasible for technical and practical reasons. However, many research labs routinely analyze for much larger numbers of PCB congeners and provide detailed congener profiles for various matrices including air and blood. This statement that more extensive analyses were not feasible for technical and practical reasons was deleted.

The statistical approach of using one-tailed significance tests and a p value of 0.10 is questionable and not well justified, and it seems especially unusual to take this approach and then apply the very conservative Bonferroni correction procedure. The authors are quite surprised about this comment because the basics of null hypothesis testing have been perfectly adapted to the present situation (in contrast to many other studies in the field):

(1) One-tailed significances are appropriate because the hypothesis is one-tailed (PCB < controls)!
(2) It is entirely appropriate to use a less conservative \( \alpha \)-level in the context of exploratory studies (in order to control for the \( \beta \)-error - since it is inappropriate not to detect subtle differences at this stage of research). This strategy is generally accepted in the behavioral sciences and requires no further justification!
(3) The reviewer may have failed to notice that the Bonferroni correction procedure, which is quite conservative, has itself been corrected according to the method of Cross and Chaffin (1982).

Thus, all relevant aspects that need to be considered for the definition of a critical significance level have been set up appropriately (and not overly conservatively) and adapted to the specific purpose of the present study. Further criticism would certainly require substantial underpinnings with appropriate statistical arguments.

Regarding the methods, what was the order of administration for the neuropsychological tests and how long did the entire battery take to administer?

The order of tests was randomised across subjects except of those memory tests that required a certain retention interval. Tests that could produce interference with the WMS long-term retention tasks were presented outside the retention interval. The administration of neuropsychological tests took about 80-90 min including a 10 min break. A brief statement was also added to the Methods section.

Given the lack of power and the relative lack of effects, there is a tendency in the discussion to over-interpret the study. We tried to reduce the potential tendency to over-interpret the results by rewriting the discussion.

A number of words, such as dopamine for example, are misspelled throughout the ms. Dopamine was corrected. The editors and reviewers are kindly requested to adapt remaining spelling mistakes during editing, please feel free to apply “close editing”.

Dopamine was corrected. The editors and reviewers are kindly requested to adapt remaining spelling mistakes during editing, please feel free to apply “close editing”.
Comments of N. Johansson:

First of all, we would like to emphasize that the present work is not a methodological contribution to PCB measurement. As explicated in the ms., the toxicology part of the current work has already been published by the state’s health authority (Gabrio et al., 2000 etc.). The relevant and new part is the neuropsychological assessment and it’s relation to internal exposure. Nevertheless, we tried to improve the sections that have been criticized by the reviewer.

1. The manuscript is too wordy and parts of it could easily be abbreviated by 50%. Several sections were abbreviated or deleted.

2. The authors mention "PCB" without indicated whether it is the sum of Dutch 6 or some other measure. The investigated PCB congeners, in particular the PCB sum value, are now more precisely described in the methods section/Biological Monitoring.

3. Please explain the meaning of d=0.7. d represents the effect size (difference of means of two measures divided by SD). A d=0.7 represents a “medium to large effect”. Effect sizes are briefly explained in the Statistical Analysis section.

4. Background: PCB congeners 105, 118 and 156 are not the most common ones, although perhaps if only mono-ortho congeners are considered. This statement was modified (see comment of other reviewers).

5. Dioxin unlike and ortho-substituted should not be equated, because a number of mono ortho-substituted congeners are regarded to be "dioxin-like". This respective text section was improved in Background.

6. The authors claim that CB 28 is ortho-substituted and dioxin-unlike. The truth is that CB 28 is mono-ortho substituted and therefore planar but regarded as not dioxin-like due to low number of chlorine atoms. We changed “nonplanar” to “low chlorinated” throughout the text. The terminology and discussion was changed as suggested by the reviewer.